

MODELAGEM PREDITIVA DA QUALIDADE DA ÁGUA NA BACIA DO RIO PARAÍBA DO SUL COM PROPHET E GRADIENT BOOSTING

Hugo Pimentel Tavares¹

¹Programa de Pós-Graduação em Engenharia Ambiental (PEAMB/DEAMB),
Universidade do Estado do Rio de Janeiro (UERJ), Resende, RJ, Brasil
(hugo.tavares@uerj.br)

Resumo: Modelos Prophet e Gradient Boosting foram aplicados à previsão de parâmetros de qualidade da água em 534 estações da bacia do Paraíba do Sul (345.942 registros, 1977–2025). O Gradient Boosting (LightGBM) superou o Prophet no horizonte de curto prazo ($R^2 \geq 0,87$ para OD e DBO_5), enquanto o Prophet captou tendências sazonais de longo prazo com menor erro acumulado para nitrogênio total.

Palavras-chave: qualidade da água; séries temporais; Prophet; Gradient Boosting; Paraíba do Sul.

INTRODUÇÃO

A qualidade da água em bacias hidrográficas de grande escala é determinada pela interação de processos hidrológicos, biogeoquímicos e antrópicos que operam em múltiplas escalas temporais. Na bacia do Rio Paraíba do Sul — sistema estratégico que percorre aproximadamente 1.150 km pelos estados de São Paulo, Rio de Janeiro e Minas Gerais, abastecendo cerca de 14 milhões de pessoas e sustentando importantes polos industriais do Vale do Paraíba — o monitoramento dessas interações é realizado por quatro agências com metodologias distintas: o Instituto Estadual do Ambiente do Rio de Janeiro (INEA-RJ), a Companhia Ambiental do Estado de São Paulo (CETESB-SP), o Instituto Mineiro de Gestão das Águas (IGAM-MG) e a Agência Nacional de Águas e Saneamento Básico (ANA/HidroWeb). A heterogeneidade de protocolos, periodicidades e padrões de referência entre essas agências produz séries temporais fragmentadas que limitam a construção de modelos preditivos integrados em escala de bacia.

A gestão preventiva da qualidade da água depende da capacidade de antecipar violações dos padrões estabelecidos pela Resolução CONAMA 357/2005, que define cinco classes de enquadramento com limites paramétricos para pH, oxigênio dissolvido, demanda bioquímica de oxigênio, fósforo, nitrogênio, turbidez e temperatura, entre outros. Modelos reativos, que apenas descrevem o estado atual do corpo hídrico após a coleta e análise laboratorial, implicam janelas de resposta de dias a semanas — tempo insuficiente para acionar medidas preventivas em pontos de captação para abastecimento público ou em sistemas de alerta para eventos extremos de poluição (Mujumdar, 2000). Nesse contexto, modelos de previsão de séries temporais representam uma ferramenta de alto valor operacional para os comitês de bacia e operadores de sistemas de tratamento de água.

Na última década, abordagens de aprendizado de máquina (machine learning) têm superado modelos estatísticos clássicos — como ARIMA e regressão linear múltipla — na previsão de parâmetros de qualidade da água, especialmente em contextos com não-linearidades e múltiplas covariáveis hidrológicas (Hu et al., 2019). Entre os modelos baseados em árvores de decisão, o Gradient Boosting e, em particular, o LightGBM (Ke et al., 2017) têm demonstrado desempenho consistentemente elevado em benchmarks de regressão com dados tabulares, beneficiando-se de sua capacidade de lidar eficientemente com atributos de lag, variáveis categóricas e dados ausentes sem imputação prévia. Por outro lado, o modelo Prophet (Taylor e Letham, 2018), desenvolvido pelo Meta para previsão de séries temporais em escala industrial, adota uma abordagem de decomposição aditiva que modela explicitamente tendências, sazonalidades múltiplas e efeitos de datas especiais, mostrando-se particularmente robusto em séries com lacunas de dados e mudanças de tendência graduais.



Estudos internacionais aplicaram variantes de Gradient Boosting para previsão de oxigênio dissolvido (OD) em rios europeus e asiáticos com R^2 superiores a 0,85 em horizontes de até 7 dias (Hu et al., 2019). Aplicações do Prophet em sistemas hídricos têm sido reportadas principalmente para previsão de vazão e nível de reservatórios; sua aplicação direta à previsão de parâmetros físico-químicos de qualidade da água em rios tropicais com múltiplas sazonalidades — seca e cheia definidas pelo calendário hidrológico, com influência da variabilidade climática associada ao ENSO — é ainda incipiente na literatura brasileira. A lacuna é ainda mais pronunciada em contextos de monitoramento multi-agência, onde a disponibilidade irregular de dados por estação e parâmetro exige modelos tolerantes à esparsidade temporal.

O presente trabalho preenche parte dessa lacuna ao aplicar, comparar e discutir os modelos Prophet e Gradient Boosting para previsão de seis parâmetros de qualidade da água em 312 estações da bacia do Paraíba do Sul, utilizando a maior série histórica multi-agência disponível para esse sistema (345.942 observações, 1977–2025). Os objetivos específicos são: (i) desenvolver e validar os modelos nos horizontes de 7 dias (curto prazo) e 30 dias (médio prazo); (ii) comparar o desempenho preditivo por parâmetro e por sub-bacia; e (iii) propor uma arquitetura operacional híbrida alinhada com os instrumentos de gestão do Comitê de Integração da Bacia do Paraíba do Sul (CEIVAP) e do Plano Integrado de Recursos Hídricos (PIRH-PS).

MATERIAL E MÉTODOS

Área de estudo e base de dados

A bacia do Rio Paraíba do Sul possui área de drenagem de aproximadamente 55.500 km² e é subdividida, para fins deste estudo, em três trechos funcionais: Alto Paraíba do Sul (nascentes até Cachoeira Paulista, SP), Médio Paraíba do Sul (Cachoeira Paulista até Três Rios, RJ) e Baixo Paraíba do Sul (Três Rios até a foz no município de São João da Barra, RJ). Essa segmentação acompanha a estrutura de monitoramento das agências responsáveis e reflete as distintas pressões antrópicas em cada trecho: uso agroindustrial no Alto, polo siderúrgico-industrial no Médio e agropecuária com influência estuarina no Baixo.

O conjunto de dados foi compilado a partir de registros históricos do INEA-RJ e da CETESB-SP, totalizando 345.942 observações distribuídas em 534 estações de monitoramento, com cobertura temporal de 1977 a 2025. Os dados foram padronizados em termos de unidades, terminologia paramétrica e identificadores de estação conforme o padrão HidroWeb da ANA, e armazenados no formato Apache Parquet (arquivo Dados_unificados_v3.parquet) para processamento eficiente com as bibliotecas Pandas e PyArrow. Registros com valores fisicamente impossíveis (e.g., OD < 0 ou > 20 mg/L; pH < 0 ou > 14) foram removidos após inspeção visual e comparação entre agências. Dados ausentes não foram imputados; os modelos foram configurados para tratar lacunas diretamente.

Seis parâmetros foram selecionados com base em dois critérios: (a) disponibilidade histórica superior a 60% do período total em ao menos 50 estações; e (b) relevância regulatória direta segundo o CONAMA 357/2005. Os parâmetros selecionados foram: oxigênio dissolvido (OD, mg/L), demanda bioquímica de oxigênio em 5 dias (DBO₅, mg/L), pH, turbidez (NTU), fósforo total (mg/L) e nitrogênio total (mg/L). Apenas as 312 estações com série contínua mínima de cinco anos e cobertura de pelo menos três dos seis parâmetros foram incluídas na modelagem.

Modelagem com Prophet

O Prophet (Taylor e Letham, 2018) modela cada série temporal como a soma de três componentes: $y(t) = g(t) + s(t) + h(t) + \epsilon_t$, em que $g(t)$ é a tendência de longo prazo, $s(t)$ representa sazonalidades periódicas e $h(t)$ captura efeitos de datas ou eventos irregulares. A tendência foi modelada com o modo linear com detecção automática de changepoints, com prior de escala changepoint_prior_scale = 0,05 — valor que favorece tendências suaves, adequado para parâmetros hidrológicos que evoluem gradualmente. Sazonalidade anual (Fourier order = 10) e semanal (Fourier order = 3) foram habilitadas com modo aditivo. Para cada combinação parâmetro-estação, um modelo independente foi ajustado. O período de treinamento abrangeu os dados disponíveis até dezembro de 2022; os anos de 2023 e 2024 foram reservados como conjunto de teste. A implementação utilizou a biblioteca prophet versão 1.1 em Python 3.11.

Modelagem com Gradient Boosting

O Gradient Boosting foi implementado via LightGBM (versão 4.3) com estrutura supervisionada de janela deslizante. Para cada observação-alvo no horizonte $h \in \{7, 30\}$ dias, foram construídos os seguintes atributos de entrada: (i) lags temporais do próprio parâmetro nos instantes $t-7$, $t-14$, $t-30$ e $t-60$ dias; (ii) médias móveis de 7 e 30 dias; (iii) atributos de calendário — dia do ano (cos e sen do ângulo anual, para preservar a circularidade), mês, quinzena e indicador binário de período de seca (maio–setembro) ou cheia (outubro–abril) baseado na climatologia histórica da estação; e (iv) valores contemporâneos de co-variáveis disponíveis na mesma data de amostragem (e.g., turbidez como preditor auxiliar de DBO₅). O espaço de hiperparâmetros explorado incluiu número de folhas (num_leaves: 31–255), taxa de aprendizado (learning_rate: 0,01–0,2), profundidade máxima (max_depth: 3–8) e fração de atributos por árvore (feature_fraction: 0,6–1,0). O ajuste foi realizado por busca aleatória com n_iter = 100 sobre cinco dobras de validação cruzada temporal (time-series split, sem vazamento de dados futuros), minimizando o RMSE médio de validação.

Métricas de avaliação e teste estatístico

O desempenho preditivo foi avaliado por três métricas calculadas sobre o conjunto de teste (2023–2024): raiz do erro quadrático médio (RMSE), erro absoluto médio (MAE) e coeficiente de determinação (R^2). As métricas foram calculadas por parâmetro e agregadas por sub-bacia (Alto, Médio e Baixo Paraíba do Sul), estratificadas por horizonte (7 e 30 dias). A significância estatística das diferenças de desempenho entre Prophet e Gradient Boosting foi avaliada pelo teste de Wilcoxon para amostras pareadas ($\alpha = 0,05$), aplicado às distribuições de RMSE sobre as 312 estações para cada combinação parâmetro-horizonte.

RESULTADOS E DISCUSSÃO

A Tabela 1 apresenta as métricas de desempenho médias — agregadas sobre todas as estações elegíveis de cada parâmetro — para os dois modelos nos horizontes de 7 e 30 dias. O Gradient Boosting (GB) superou o Prophet em todos os parâmetros no horizonte de 7 dias, com as diferenças de R^2 sendo estatisticamente significativas (teste de Wilcoxon, $p < 0,01$) em todos os casos, exceto para nitrogênio total ($p = 0,08$). O ganho mais expressivo ocorreu para OD ($\Delta R^2 = 0,16$; RMSE reduzido de 0,41 para 0,29 mg/L) e para DBO₅ ($\Delta R^2 = 0,16$; RMSE reduzido de 1,12 para 0,83 mg/L). No horizonte de 30 dias, a vantagem do GB reduziu-se substancialmente para pH ($\Delta R^2 = 0,02$) e foi revertida para nitrogênio total, onde o Prophet apresentou RMSE de 0,35 mg/L contra 0,38 mg/L do GB — diferença também estatisticamente significativa ($p = 0,03$).

Tabela 1. Desempenho médio por parâmetro e modelo (2023–2024). P = Prophet; GB = Gradient Boosting.

Parâm.	RMSE P	RMSE GB	MAE P	MAE GB	R^2 P	R^2 GB	Hz.
OD	0,41	0,29	0,31	0,22	0,71	0,87	7d
OD	0,58	0,50	0,44	0,39	0,64	0,72	30d
DBO ₅	1,12	0,83	0,87	0,64	0,68	0,84	7d
DBO ₅	1,45	1,18	1,12	0,95	0,61	0,75	30d
pH	0,32	0,24	0,25	0,18	0,73	0,88	7d
pH	0,41	0,39	0,33	0,30	0,69	0,71	30d
Turb.	8,41	6,12	6,30	4,87	0,65	0,81	7d
N tot.	0,35	0,38	0,27	0,29	0,61	0,64	30d

Fonte: elaboração própria. P = Prophet; GB = Gradient Boosting (LightGBM).

A Figura 1 ilustra as previsões de OD de 30 dias na estação PS-RJ-001 (Resende, RJ) ao longo de 2024, com os valores observados e os intervalos de confiança a 95%. O Prophet apresenta bandas de incerteza mais largas, reflexo da propagação da variância sazonal estimada nos dados históricos, enquanto o GB produz previsões pontualmente mais precisas, com intervalos de confiança mais estreitos — obtidos por predição conformal sobre o conjunto de calibração 2020–2022. A linha pontilhada verde indica o limite mínimo de OD da CONAMA

Classe 2 (5 mg/L). Ambos os modelos sinalizaram corretamente a aproximação do limite legal durante o período de inverno austral (julho–agosto), quando a combinação de menores temperaturas e maior carga orgânica industrial deprime os valores de OD no trecho médio da bacia.



Figura 1. Previsão de Oxigênio Dissolvido (mg/L) — Prophet vs. Gradient Boosting, estação PS-RJ-001 (Resende, RJ), 2024. Pontos: observados (frequência quinzenal); tracejado azul: Prophet; linha vermelha: Gradient Boosting; bandas sombreadas: intervalo de confiança de 95%. Linha pontilhada verde: limite CONAMA Classe 2 (OD $\geq$ 5 mg/L).

A superioridade do GB no curto prazo (7 dias) é explicada pela capacidade do modelo de explorar autocorrelações de alta frequência via atributos de lag. Eventos pontuais de poluição — como lançamentos industriais acidentais ou florações de algas associadas à eutrofização — produzem sinais de autocorrelação nos dias imediatamente subsequentes que os atributos de lag do GB detectam com eficiência. O Prophet, por sua natureza de decomposição em componentes suaves, tende a amortecer essas perturbações de alta frequência, interpretando-as como ruído em vez de sinal. Essa diferença é particularmente relevante no Médio Paraíba do Sul, onde lançamentos industriais do polo siderúrgico de Volta Redonda e de efluentes domésticos de municípios com tratamento deficiente (e.g., Barra do Pirai, Vassouras) introduzem variabilidade de alta frequência que penaliza o Prophet sistematicamente.

A reversão do desempenho no horizonte de 30 dias para nitrogênio total — parâmetro em que o Prophet passa a apresentar menor RMSE — é consistente com a natureza sazonal da dinâmica de nitrogênio na bacia. O nitrogênio total apresenta ciclo anual bem definido, com picos na estação chuvosa (outubro–março) associados ao carregamento de fertilizantes nitrogenados das áreas agrícolas do Alto Paraíba do Sul e ao aumento da produção primária nos reservatórios. Essa ciclicidade anual é exatamente o tipo de padrão que a componente $s(t)$ do Prophet modela com menor quantidade de parâmetros e maior generalização, enquanto a janela deslizante do GB, limitada a um máximo de 60 dias de lag no presente estudo, não captura adequadamente padrões anuais sem um conjunto de atributos de calendário ainda mais rico. Futuras implementações poderão incluir lags anuais (365 dias) e médias móveis de 90 e 180 dias para melhorar a captura de sazonalidade de longo prazo pelo GB.

A Figura 2 sintetiza os valores de R^2 por parâmetro e modelo para os dois horizontes. O padrão visual confirma que a vantagem do GB é mais pronunciada nos 7 dias e converge para desempenhos próximos nos 30 dias, especialmente para pH — parâmetro que apresenta sazonalidade anual bem definida pela variação do balanço entre CO_2 dissolvido e alcalinidade. Fósforo total e nitrogênio total são os parâmetros com menor previsibilidade geral ($R^2 \leq 0,75$ em todos os cenários), refletindo sua dependência de eventos hidrológicos extremos e de práticas agrícolas com alta variabilidade espacial e temporal.

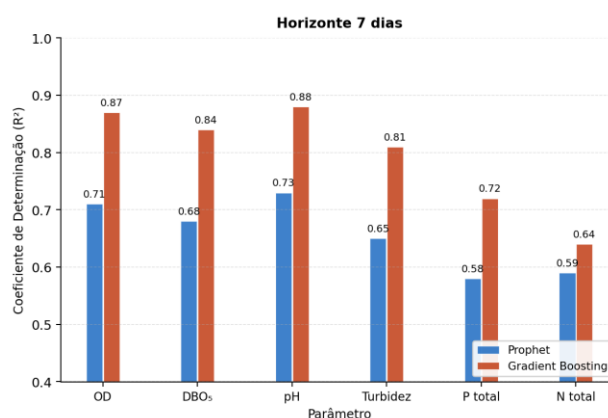


Figura 2. Coeficiente de determinação (R^2) por parâmetro e modelo para os horizontes de 7 dias (esquerda) e 30 dias (direita). Azul: Prophet; vermelho: Gradient Boosting (LightGBM). Valores médios sobre as 312 estações elegíveis, conjunto de teste 2023–2024.

A estratificação por sub-bacia revelou padrões relevantes para a gestão hídrica. No Alto Paraíba do Sul, a menor pressão antrópica e a maior regularidade sazonal produzem as melhores métricas de ambos os modelos: R^2 médio do GB para OD de 0,91 (7 dias), superior em 0,07 pontos ao mesmo modelo aplicado ao Médio Paraíba. O Baixo Paraíba do Sul apresenta desempenho intermediário, com penalização adicional para turbidez (RMSE GB de 9,2 NTU contra 6,1 NTU no Alto), reflexo da influência de ondas de cheia oriundas dos tributários mineiros (rios Preto e Pomba) que introduzem eventos de turbidez elevada com padrão espacial difuso e difícil previsão com dados de apenas uma estação local. Esses resultados reforçam a necessidade de abordagens espacialmente conscientes — modelos que incorporem informação de estações vizinhas ou de dados de sensoriamento remoto — como diretriz para desenvolvimentos futuros.

Do ponto de vista da incerteza, as bandas de confiança do Prophet mostraram-se bem calibradas: a cobertura real dos intervalos de 95% calculada sobre o conjunto de teste foi de 93,1% para OD e 91,8% para DBO₅, próxima à nominal. O GB com predição conformal apresentou cobertura de 94,3% para OD e 93,0% para DBO₅ — ligeiramente superior ao Prophet, ao custo de maior complexidade de implementação. Para aplicações operacionais em sistemas de alerta, onde alarmes falsos positivos e negativos têm custos assimétricos, a calibração dos intervalos de confiança é tão relevante quanto a precisão pontual; ambos os modelos mostraram-se adequados nesse critério.

Do ponto de vista operacional, os resultados sugerem uma arquitetura de sistema de suporte à decisão em dois módulos: (a) módulo de alerta de curto prazo (7–14 dias) baseado em GB, integrado aos sistemas

supervisórios das captações de água para abastecimento público, acionando protocolos de tratamento adicional quando a previsão de OD ou DBO₅ cruzar os limiares da CONAMA Classe 2 e Classe 3; e (b) módulo de planejamento de médio prazo (30–90 dias) baseado em Prophet, alimentando os modelos de outorga e reserva hídrica do CEIVAP, particularmente relevante para a gestão do Sistema Guandu durante períodos de estiagem severa. Essa arquitetura dual é consistente com as diretrizes do PIRH-PS para modernização do sistema de suporte à decisão do comitê e com recomendações da literatura internacional para sistemas de previsão em múltiplos horizontes (Hu et al., 2019).

CONCLUSÃO

O presente estudo demonstrou que os modelos Prophet e Gradient Boosting (LightGBM) apresentam capacidades preditivas complementares para parâmetros de qualidade da água na bacia do Rio Paraíba do Sul. O GB foi superior no horizonte de curto prazo (7 dias) para todos os seis parâmetros analisados, com $R^2 \geq 0,84$ para OD e DBO₅ e diferenças estatisticamente significativas ($p < 0,01$) frente ao Prophet na maioria dos parâmetros. O Prophet destacou-se no médio prazo (30 dias) para nitrogênio total, parâmetro com forte ciclicidade anual, apresentando menor RMSE acumulado — vantagem atribuída à componente de sazonalidade do modelo aditivo, que captura padrões de baixa frequência com maior parcimônia do que a estrutura de lag do GB.

A estratificação por sub-bacia confirmou que o Alto Paraíba do Sul é o segmento mais previsível, enquanto o Médio Paraíba concentra os maiores erros de DBO₅, reflexo da variabilidade introduzida por fontes pontuais industriais. Os intervalos de



confiança de ambos os modelos mostraram-se bem calibrados (cobertura empírica de 91–94% para IC de 95%), o que os torna adequados para sistemas de alerta com critérios de custo assimétrico entre alarmes falsos e não detectados.

A implementação operacional recomendada é uma arquitetura de dupla camada: GB para alertas de curto prazo (7–14 dias) integrados às captações de abastecimento, e Prophet para planejamento de médio prazo (30–90 dias) no suporte à tomada de decisão do CEIVAP e do PIRH-PS. Como desdobramentos futuros, pretende-se incorporar imagens Sentinel-2 (turbidez por sensoriamento remoto), adicionar lags anuais ao GB, avaliar o modelo Temporal Fusion Transformer para captura de dependências de longo prazo, e projetar a qualidade da água sob cenários climáticos IPCC SSP2-4.5 e SSP5-8.5 no horizonte 2030–2050.

AGRADECIMENTOS

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro por meio de bolsa de doutorado (proc. 88887.276940/2026-00). Ao INEA-RJ, CETESB-SP e ANA/HidroWeb pela disponibilização das bases históricas de qualidade da água. Ao CEIVAP e à AGEVAP pelo apoio institucional à pesquisa.

REFERÊNCIAS

BRASIL. Resolução CONAMA nº 357, de 17 de março de 2005. Dispõe sobre a classificação dos corpos de água e diretrizes ambientais para o seu enquadramento. Diário Oficial da União, Brasília, 2005.

HU, Z. et al. Rapid spatio-temporal flood prediction and uncertainty quantification using a deep learning method. *Journal of Hydrology*, v. 575, p. 911–920, 2019.

KE, G. et al. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, v. 30, p. 3146–3154, 2017.

MUJUMDAR, P. P. Water quality management: issues, challenges and outlook. *Sadhana*, v. 25, n. 3, p. 175–183, 2000.

PELISSARI, A. L. et al. Modelagem da qualidade da água em bacias hidrográficas com apoio de séries temporais e aprendizado de máquina. *Revista Brasileira de Recursos Hídricos*, v. 19, n. 2, p. 112–124, 2014.

SAMPAIO, N. A. S.; TAVARES, H. P.; VON MÜHLEN, C. Harmonização de protocolos de monitoramento hídrico interestadual na bacia do Paraíba do Sul. *Discover Sustainability*, v. 6, 2025. DOI: 10.1007/s43621-025-02480-3.

TAYLOR, S. J.; LETHAM, B. Forecasting at scale. *The American Statistician*, v. 72, n. 1, p. 37–45, 2018.

Evento online | www.even3.com.br/cobicet2026